Perceptual Segmentation and Component Selection in Compact Sinusoidal Representations of Audio+

Ted Painter
Intel Corporation HD2-230
Handheld Computing Division
77 Reed Road Hudson, MA 01749

Andreas Spanias
Department of Electrical Engineering
Arizona State University,
Tempe, AZ 85287-7206, U.S.A.

Abstract

This paper presents two fundamental enhancements in a hybrid audio signal model consisting of sinusoidal, transient, and noise (STN) components. The first enhancement involves a novel application of a perceptual metric for optimal time segmentation for the analysis of transients. In particular, Moore and Glasberg's model of partial loudness is modified for use with general signals and then integrated into a novel time segmentation scheme. The second and perhaps more significant STN enhancement is concerned with a new methodology for ranking and selection of the most perceptually relevant sinusoids.

1. Introduction

Signal-adaptive modeling has recently been adopted in sinusoidal modeling algorithms [1,2,3]. Existing work [3,4] has addressed multi-resolution analysis. The enhancement to the existing STN model proposed in this paper attempts to eliminate computationally expensive and bit allocation intensive filter banks from the STN model. These hybrid structures are replaced with combined use of sinusoidal analysis and perceptually-controlled time segmentation scheme that activates short analysis windows only during transient events that are judged to be unmasked.

The paper essentially presents two new methods for perceptual segmentation and selection of sinusoids in hybrid (STN) sinusoidal modeling of audio. The first contribution of this paper is that it adopts a perceptual metric for optimal time segmentation for the analysis of transients. The second and perhaps more significant STN contribution deals with a new methodology for ranking and selection of the most perceptually relevant sinusoids. The idea behind the method, known as Excitation Similarity Weighting (ESW), is to maximize the matching between the auditory excitation pattern associated with the original signal and the corresponding auditory excitation pattern associated with the modeled signal that is being represented by only a few sinusoidal parameters. The reconstruction quality provided by ESW is compared against a quality benchmark associated with the maximum signal-to-mask ratio (maximum SMR) methodology. The ESW component selection methodology is shown to outperform the maximum SMR selection strategy in terms of both objective and subjective quality.

The paper is organized as follows. First, a review of the classical sinusoidal model is given in section 2. Section 2.2 presents the STN extensions to the basic sinusoidal model and verification results for the partial loudness adaptation metric along with an enhanced partial loudness time segmentation scheme for processing of transients. Finally, section 2.3 describes strategies for STN model pruning and the ESW sinusoidal component selection methodology. Sample results are given for application of the method to a spectrally complex signal.

2. The Hybrid Adaptive Sinusoidal Model: Sines + Transients + Noise (STN)

The classical sinusoidal model comprises an analysis-synthesis framework [5] that represents a signal, $s(n)$, as the sum of a collection of K sinusoids (*"partials"*) with time-varying frequencies, phases, and amplitudes, i.e.,

$$s(n) = \hat{s}(n) = \sum_{k=1}^{K} A_k \cos(\omega_k(n)n + \phi_k(n)) \quad (1)$$

where $A_k(n)$ represents the amplitude, $\omega_k(n)$ represents the instantaneous frequency, and $\phi_k(n)$ represents the instantaneous phase of the k^{th} sinusoid. Estimation of parameters is typically accomplished by peak picking the short-time Fourier transform (STFT) [5]. In the synthesis stage, the model parameters are subjected to spectral line tracking and frame-to-frame amplitude and phase interpolation.

Although the basic sinusoidal model achieves efficient representation of harmonically structured signals, extensions to the basic model have also been proposed for signals containing non-tonal energy [6] The spectral modeling and synthesis system (SMS) treats audio as the sum of K sinusoids along with a stochastic component, $e(n)$, i.e.,

$$s(n) = \hat{s}(n) = \sum_{k=1}^{K} A_k \cos(\omega_k(n)n + \phi_k(n)) + e(n) \quad (2)$$

Although the sines + noise signal model gave improved performance, the addition of transient components giving rise to a three-part model consisting of sines + transients + noise (STN) [4,7] (Fig. 1) provides additional enhancements. In fact, the focus and the contribution of this paper is on

+ Research performed at Arizona State University as part of a Ph.D. thesis project of Dr. Painter (Advisor: A. Spanias)

0-7803-7041-4/01/$10.00 ©2001 IEEE

optimization of the STN model for a scalable audio coding application [8].

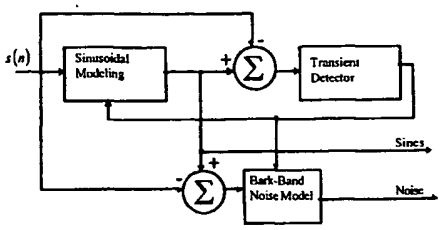

Fig. 1. Sines+Transients+Noise Model (STN).

2.1 STN Adaptive Time/Frequency Analysis

Signal-adaptive modeling has recently been adopted in sinusoidal analysis-synthesis algorithms [3,4]. The enhancement to the existing STN model [4] proposed in this paper is a perceptually controlled time segmentation scheme that activates short analysis windows only during transient events that are judged to be unmasked. The transients are detected via an energy threshold combined with a partial loudness edge detection scheme that operates on the sinusoidal modeling residual. Both masked and unmasked transients can potentially trip the energy threshold detector, but masked transients have a significantly lower impact on residual noise loudness than unmasked transients. The advantage of the proposed system is that it avoids overly conservative coding of masked transients.

2.2 STN Transient Processing and Partial Loudness Metric Verification

The proposed transient detection scheme uses Moore and Glasberg's Partial Loudness (PL) model to measure the partial loudness of the modeling residual in each frame. Perceptually relevant (unmasked) transients are detected when the partial loudness (in Sones) exceeds the mean loudness of previous frames by a factor of 2.0 to 2.7. Because it is necessary to make modifications to the partial loudness model for audio signals, a series of verification tests were conducted and masking and loudness predictions were compared against well-established psychophysical listening tests. The modified model predictions were tested for noise-masking-tone experiments in which a series of tones were masked by a 90 Hz narrowband masker centered at 410 Hz, with the masker presented at levels of 40, 60, and 80 dB SPL. Model predictions were compared against the data reported by Egan and Hake [9]. A noise loudness of 0.003 Sone was assumed to correspond to the masked threshold for the probe tones. In all cases the model predictions were consistent with the experimental data. A sample result appears in Fig. 2. A second set of experiments was conducted to measure equal loudness contours and minimum audible field (MAF) predictions.

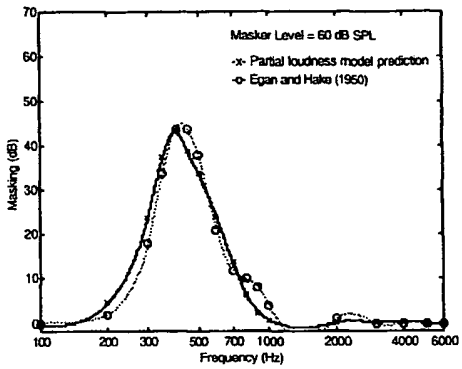

Fig. 2. Masking predictions on noise masking experiment for 60 dB SPL 90 Hz bandwidth noise masker centered at 410 Hz.

The results compared favorably against ISO 389-7:1996(E). The performance trends exhibited in these tests as well as in informal listening tests with more complex auditory stimuli instilled confidence in the metric to the extent that it was judged to be a good predictor of partial loudness. A perceptually salient transient processing scheme was devised with this PL metric as the core processing block. Additional simulations [8] demonstrated the complete processing chain of the proposed transient detection and pre-echo compensation scheme. It was shown that the transient processor is able to identify the most perceptually significant transients, and that the scheme provides a natural scaling mechanism for low rate applications. Transients are ranked in terms of partial loudness, and then bits are allocated to the transient events in order of significance. The results of informal listening tests confirmed that the method provided a valid perceptual ranking of transient salience. After transient processing, the sinusoidal+transient modeling residual is captured using an FFT-based Bark-band structure [4].

2.3 Compact Representation of STN Parameters

The second and most significant STN enhancement proposed in this paper is concerned with the ranking and selection of perceptually relevant sinusoids on a compact set. We call this the Excitation Similarity Weighting (ESW) ranking and selection procedure. Whereas current coders tend to choose maximum signal-to-mask ratio components and therefore base the selection decision on the masked threshold associated with the original signal, the ESW methodology seeks to maximize the matching between the excitation patterns evoked by the coded and original signals on a short-time basis. In this way, ESW does not seek to satisfy noise threshold criteria. In contrast to ESW, the maximum SMR selection criterion does not guarantee maximal matching between the modeled and the original excitation patterns [8]. The idea behind the ESW technique

is to select sinusoids in such a way that each new sinusoid added to a modeled representation is guaranteed to provide a maximum incremental gain in matching between the auditory excitation pattern associated with the original signal and the corresponding auditory excitation pattern associated with the modeled signal. In order to accomplish this goal, an iterative process is proposed in which each sinusoid extracted during conventional analysis is assigned an excitation similarity weight. During each iteration, the sinusoid having the largest weight is added to the modeled representation. New sinusoids are accumulated until some constraint is exhausted, e.g., a bit budget. The algorithm tends to converge as the number of modeled sinusoids increases. The ESW sinusoidal component selection strategy (Fig. 3) works as follows. First, a complete set of sinusoids is estimated using the STFT. Then, a reference excitation pattern is computed for the original signal in a manner similar to the method outlined in the description of PERCEVAL [10]. The pattern may contain up to 2500 discrete excitation levels that correspond to assumed discrete detectors along the basilar membrane. An iterative ranking procedure is performed next. The objective on the k-th iteration is to extract from the candidate set the most perceptually salient sinusoid, given the previous k-1 selections. The method assumes that maximum perceptual salience is associated with the component able to affect the greatest improvement in matching between the excitation pattern associated with the original signal and the excitation pattern that is associated with the modeled signal. To select from among the candidates during the k-th iteration, a complete set of candidate excitation patterns is computed, one each for the patterns associated with the modeled signal containing the first k-1 selected sinusoids, as well as each of the candidates currently available. The candidate that minimizes the difference between the reference and modeled excitation patterns is selected for the k-th iteration. The resulting sinusoidal parameters of the best candidate are passed to the trajectory tracking and model pruning components. The core ESW calculation comprises an average difference calculation that operates on the reference and test excitation patterns. In particular, the average difference, Δ_k, between the original (reference) and the test patterns on the k^{th} iteration is given by

$$\Delta_k = \frac{1}{D}\sum_{i=1}^{D} E(i) - X_k(i) \quad (3)$$

where $E(i)$ is the reference excitation pattern level (in dB), $X_k(i)$ is the level (in dB) any of the candidate test excitation patterns on the k^{th} iteration, and D is the number of detectors. Therefore, for each pattern, the improvement in matching on the k^{th} iteration for each candidate pattern, $X_k(i)$, is given by

$$\Delta_k - \Delta_{k+1} = \frac{1}{D}\sum_{i=1}^{D} X_{k+1}(i) - X_k(i) \quad (4)$$

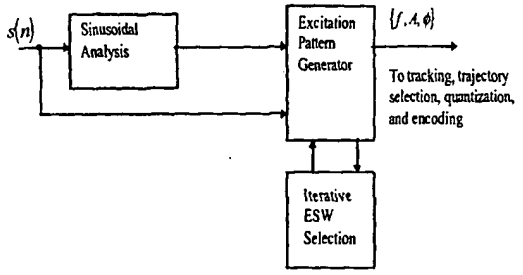

Fig. 3. The ESW scheme.

The ESW technique computes the matching improvement for all candidate patterns during the k^{th} iteration and selects the component that maximizes Eq. 4. Once the best candidate pattern, $X_k^*(i)$, has been identified on the k^{th} iteration (in the sense of maximizing Eq. (4), an excitation similarity weight is assigned to the sinusoidal component that provided the maximum incremental matching improvement. The ESW assigned to the k^{th} component is

$$ESW_k = \Delta_{k-1} - \Delta_k \quad (5)$$

2.4 Comparison of ESW Versus Maximum SMR

For validation, the ESW component selection and ranking scheme was compared against a reference maximum-SMR selection scheme over a diverse collection of audio program material. The ESW based output samples generated from STN model parameters consistently outperformed the SMR based audio samples in terms of both subjective informal listening tests and objective evaluations using the partial loudness model described earlier. We give here sample comparative results in graphical format for a selection of rock music that was judged to be spectrally complex and therefore challenging for a low rate coding application. The pair of figures shown provides insight on how the ESW methodology selects components in contrast to the maximum SMR methodology. These comparative results (Fig. 4) show a spectral view corresponding to 23 milliseconds of audio. The vertical arrows in both figure panels correspond to the complete set of sinusoids returned by classical sinusoidal analysis. The dashed line corresponds to a short-time spectral estimate (magnitude FFT) mapped to SPL, and the solid line corresponds to an estimate of the masked threshold generated by the MPEG-1 psychoacoustic model 2. Sinusoids labeled in panel (a) of the figure were selected on the basis of maximum SMR. Each of the selected sinusoids is labeled with its rank, one through ten, and its SMR, in dB. It is clear from the figure that the ranking is in terms of

descending SMR. This ranking corresponds directly to the currently popular method of sinusoid selection. Panel (b) of the figure shows the selection process for the ESW methodology. In this figure, each of the ten selected sinusoids is labeled with its rank and ESW score (Eq. 5). A comparison of the figures reveals that the ESW method tends to choose sinusoids across the spectrum, whereas the maximum SMR method tends to choose sinusoids of higher energy that are clustered at lower frequencies. This trend was manifested across time in the given example and also across many musical selections.

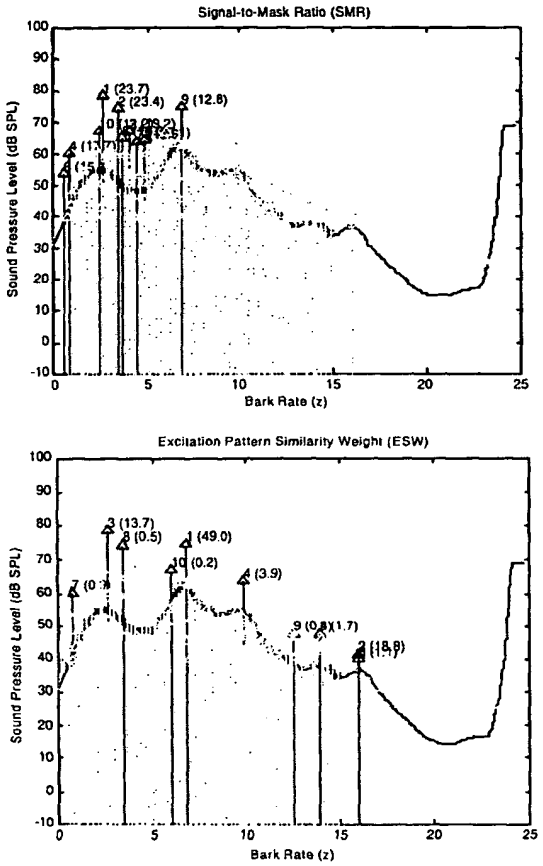

Fig. 4. Comparison of sinusoidal component pruning methodologies: (a) Maximum SMR selection; (b) Maximum ESW selection.

3. Concluding Remarks

The enhancements in sinusoidal selection have been shown to lead to several methods for achieving compact representations of ESW-ranked sinusoidal components. Perhaps the most intuitive is that of thresholding on the basis of a minimum ESW. All sinusoids below the minimum ESW can be discarded. Because ESW ranks components in order of their impact on excitation pattern matching, the threshold can be used as a rate or quality scaling factor. Experimentation with trajectory ranking by individual component as well as minimum, mean and maximum ESW over time has also shown significant promise in informal listening tests for pruning the set of trajectories that are generated during STN analysis. Pruning mean ESW components below a certain threshold has in some cases reduced the number of trajectories by 15 to 20% without audible impact.

REFERENCES

[1] D. V. Anderson, "Speech Analysis and Coding Using a Multi-Resolution Sinusoidal Transform," in *Proc. ICASSP-96*, pp. 1045-1048, May 1996.

[2] D. Ellis and B. Vercoe, "A Wavelet-Based Sinusoid Model of Sound for Auditory Signal Separation," *in Proc. Int. Comp. Mus. Conf.*, pp. 86-89, 1991.

[3] M. Goodwin, "Multiresolution Sinusoidal Modeling Using Adaptive Segmentation," *ICASSP-98*, May 1998

[4] S. Levine and J. Smith, "A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications," in *Proc. 105th Conv. Aud. Eng. Soc.*, preprint #4781, Sep. 1998.

[5] R. McAulay and T. Quatieri, "Speech Analysis Synthesis Based on a Sinusoidal Representation," *IEEE Trans. ASSP*, pp. 744-754, Aug. 1986.

[6] X. Serra, *A System for Sound Analysis/Transformation /Synthesis Based on A Deterministic Plus Stochastic Decomposition.* Ph.D. Thesis, Stanford University, 1989.

[7] T. Verma, *et al.*, "Transient Modeling Synthesis: A Flexible Analysis/Synthesis Tool for Transient Signals," in *Proc. Int. Comp. Mus. Conf.*, 1997.

[8] T. Painter, "Scalable Perceptual Audio Coding withA Hybrid Adaptive Sinusoidal Signal Model,", Ph.D. Thesis, Arizona State University, June 2000.

[9] J. Egan and H. Hake, "On the Masking Pattern of a Simple Auditory Stimulus," *J. Acoust. Soc. Am.*, vol. 22, pp. 622-630, 1950.

[10] B. Paillard, *et al.*, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Aud. Eng. Soc.*, v. 40, n. 1/2, pp. 21-31, Jan./Feb. 1992.